Evaluating Causal-Based Fairness Metrics and Statistical Fairness Metrics

Adam Lin adam.lin@columbia.edu

Andrea Clark aoc2111@columbia.edu

Rohith Ravindranath rr3415@columbia.edu

Abstract

This work aims to explore established causal-based frameworks and evaluate them under causality-based fairness notions, comparing these measures to statistical fairness metrics. A major limitation of existing causal literature is that it assumes that a causal model to work from, which is rarely the case. For this reason, part of our work will be analyzed ways to discover such a causal model through which we will evaluate the fairness notions specifically the direct and indirect effect of the sensitive attribute on the outcome. In this paper, we developed a casual fairness pipeline for observational data. This pipeline can be applied to analyze classification outcomes and give insight into the effect of statistical fairness mitigation algorithms.

1. Introduction

Addressing the notion of fairness in machine learning has become a more ubiquitous problem in recent years, as learning algorithms have been more involved in making societal decisions in various aspects, including job hiring, college admission, and loan applications, to name a few. Traditional fairness metrics, such as demographic parity and equalized odds, strictly depend on the joint distribution of the observational criteria and have been shown to fail to detect unfairness in the presence of statistical anomalies, such as Simpson's paradox Simpson (1951).

Examining fairness through the lens of causality has been recently gaining traction as the accepted method for appropriately addressing the notion of fairness, as statistical fairness notions face many challenges, such as subgroups yielding different fairness results when compared with the entire population. An example of this phenomenon is evidenced by the IMPACT teacher evaluation system, which is claimed to be biased against teachers who have been assigned to students at a lower starting academic level, where the income level of the location of the school influences the level of the students in the classO'Neil (2016). Fairness in this situation can only be claimed when the dependence between confounder and protected attribute is broken, that is, the firing rates of teachers are examined through an intervention on the starting level of students in the teacher's class: are firing rates equal when *all* teachers are assigned high-level students versus low-level students? This real-world example illustrates the need to consider causal frameworks when addressing fairness,

as traditional statistical metrics based on correlation would have missed the implicit bias in the data-generating process: very few teachers in high-income schools were found to be assigned to low-level students Makhlouf et al. (2020b).

This work aims to further the analysis of causal fairness at the practical level. While much work has been done framing the notion of fairness under a causal framework, such as individual fairness Dwork et al. (2012) Joseph et al. (2016) Louizos et al. (2016) Zemel et al. (2013), fairness through unawareness Grgic-Hlaca et al. (2016), counterfactual fairness Kusner et al. (2017), we seek to establish a causal model from which to baseline causal-based fairness metrics and compare these results to statistical fairness metrics.

2. Related Work

2.1 Statistical-based fairness metrics

Most statistical-based fairness metrics can be separated into three different categories: independence, separation, and sufficiency Barocas et al. (2019). Given a sensitive attribute A, independence fairness metrics are satisfied when $\hat{Y} \perp A$. This means that the proportion of the predicted outcome should be the same between different groups. Notice that this does not allow A to be a proxy for $\hat{Y}$. Separation states that A can serve as a proxy for $\hat{Y}$ as long as $\hat{Y}$, but should be independent to A given Y ($\hat{Y} \perp A|Y$). Satisfying sufficiency ensures that a classifier is well-calibrated for each group of the sensitive attribute ($Y \perp A|\hat{Y}$).

One of fairness metrics that falls under the category of independence is **statistical parity**, also known as **demographic parity**, and is one of the most commonly accepted fairness metrics Dwork et al. (2012). A classifier $\hat{Y}$ satisfies statistical parity if:

$$\Pr(\hat{Y} \mid A = 0) = \Pr(\hat{Y} \mid A = 1)$$

This metric requires the prediction to be conditionally independent of the sensitive attribute. The main drawback of this metric is that when base rates of the label are unequal when looking at the population partitioned by the sensitive attribute, this metric can be misleading. Another drawback is the potential for what Barocas and Selbst refer to as "masking", in which the model optimizes performance for the majority group and negatively impacts another protected group by random-selection to achieve the fairness notion, which is particularly a problem for statistical parity Makhlouf et al. (2021).

Unlike statistical parity, **equalized odds**, which falls under the category of separation, takes into account both the predicted and actual outcomes and allows $\hat{Y}$ to depend on A but only through the target variable Y Hardt et al. (2016). Formally,

$$\Pr(\hat{Y} = 1 \mid Y = y,\ A = 0) = \Pr(\hat{Y} = 1 \mid Y = y,\ A = 1)\ \forall y \in \{0, 1\}$$

For the outcome $y = 1$, this implies that $\hat{Y}$ must achieve an equal true positive rate (TPR) across groups, and for $y = 0$, $\hat{Y}$ must achieve an equal false positive rate (FPR) across groups. This mitigates the "laziness" problem present in statistical parity by punishing models that optimize accuracy for the majority group Makhlouf et al. (2021). We will focus mainly on these two metrics in our experiments.

2.2 Casual based fairness

We now turn our focus to causal-based fairness, as it is the core of this work. The most common non-causal fairness notion is total variation (TV) - such as statistical parity, de-

mographic parity, or risk difference. One of the biggest limitations with respect to TV is that it is purely statistical nature which makes it unable to to reflect the causal relationship between the sensitive attribute and the outcome. With this in mind, causal-based fairness can give another perspective (Makhlouf et al. (2020a)).

Within causal-based fairness, there are two different frameworks - disparate impact (Barocas and Selbst (2016); Plecko and Bareinboim (2022)) and disparate treatment (Barocas and Selbst (2016)). Disparate impact aims at ensuring the equality of outcomes across all groups. Disparate treatment seeks equality of treatment achievable through prohibiting the use of the sensitive attribute in the decision process.

One example of a causal-based fairness metric in the disparate impact framework is the total effect (TE) (Makhlouf et al. (2020a)). TE is the causal version of TV and is defined in terms of experimental probabilities as follows:

$$TE_{a_1,a_0}(y) = \Pr(y_{a_1}) - \Pr(y_{a_0})$$

where $A = a_0$ denotes the privileged group and $A = a_1$ the disadvantaged group. TE measures the effect of the change of A from a_1 to a_0 on $Y = y$ along all the causal paths from A to Y. We also remark that while TV reflects the difference in proportions of $Y = y$ in the current cohort, TE reflects the difference in proportions of $Y = y$ in the entire population.

With respect to the disparate treatment framework, the common causal-based fairness metrics include direct effect, indirect effect and path-specific effect (Pearl (2001)). An effect can be deemed fair, unfair, or spurious by an expert of the scenario at hand. The specific notion for direct effect (DE) is

$$DE_{a_1,a_0}(y) = \Pr(y_{a_1}, Z_{a_0}) - \Pr(y_{a_0})$$

Now, here Z is the set of mediators between paths from the sensitive attribute to the target label Y ($A \to Z \to Y$). By assuming that all Z take values $A = a_0$, we can measure the effect from when $A = a_1$, essentially masking off all of the indirect effects. Similarly, the indirect effect (IE) works by masking the effect of the direct path and assuming all the mediators Z take values the natural value of when $A = a_1$.

$$IE_{a_1,a_0}(y) = \Pr(y_{a_0}, Z_{a_1}) - \Pr(y_{a_0})$$

Indirect effect is assessed using the causal effect along the paths that pass through proxy attributes. A high value of direct effect implies that there is some sort of discrimination, but indirect effect can be decomposed into those effects that are explainable and those that are true discrimination. A fair or explainable discrimination is measured using causal pathways passing through explaining variables, which we can denote as $ED(Y)$. We also denote the discriminating path effect as indirect discrimination ($ID(Y)$). Each of these causal effects can be estimated through observational data Zhou and Yamamoto (2020).

3. Methods

3.1 Datasets

3.1.1 ACSIncome

The primary dataset used for this analysis was the ACSIncome, compiled by Ding et al. (2021) as an alternative to the UCI Adult dataset Kohavi and Becker (1996). It was

reconstructed with the IPUMS interface to the Current Population Survey (CPS) data from 1994 Ruggles (2021). The prediction task for this dataset mirrors that of the UCI Adult: predicting whether an individual's income is above $50,000. It contains 1,599,229 observations and 10 different features, of which we used 7 features, following the causal structure modeling of the UCI Adult dataset as per Binkyt.e-Sadauskien.e et al. (2022). The summary of the features used in our modeling is listed in Table 2 in Appendix A.

3.1.2 UCI Adult

The dataset we used for baseline comparison is the well-known UCI Adult dataset Kohavi and Becker (1996), compiled from the 1994 Census database. As a model for the newer ACSIncome dataset, it naturally also predicts whether or not an individual earns above $50,000 annually. It contains 48,842 observations and 14 features, a subset of which are identical to those present in the ACSIncome dataset. For our modeling, we used a subset of features of the UCI Adult dataset that were also present in the ACSIncome dataset.

3.2 Pipeline

The statistical framework is limited in that it cannot address notions of bias that are introduced through proxy features that may be correlated with the sensitive attribute(s). In this scenario, the causal framework has an additional branch that explains this potential mechanism of bias, which we denote as an *indirect effect*. The indirect effect can take on two forms: *explainable discrimination* and *indirect discrimination*. The former is legitimate form of bias that is uncoupled from the sensitive attribute, while the latter allows for bias to unfairly influence the target through a proxy feature for the sensitive attribute.

To draw a fair comparison between the statistical and causal fairness metrics, we first develop a pipeline to calculate the casual fairness values. As shown in Fig 5, our pipeline will consist of first learning a causal structure model (CSM) for our chosen datasets. For our experiments, we just use the PC algorithm for structure learning, as our focus is on evaluating the sensitivity of causal structure metrics on addressing bias and drawing a comparison to statistical causal metrics (Figure 5).

After learning the CSM, we use this to first measure the direct and indirect effects on the raw data and compare these causal metrics to the equal opportunity and disparate impact statistical metrics. As the goal of addressing fairness is to ultimately correct bias present in the data, we then perform four different bias correcting algorithms to see how sensitive each metric is to shifts in the data: adversary debiasing Lemoine et al. (2018), reweighing (Kamiran and Calders (2012)), rejection option (Kamiran et al. (2012)), and optimized preprocessing (Calmon et al. (2017)).

3.3 Linking statistical fairness metrics and casual fairness metrics

Typically, causal fairness measure is calculated on the true label Y. We can answer question of if there is bias in the data generation process by calculating the direct and indirect effect of A on Y. Now, if causal effect is calculated on a predicted outcome of a classifier $\hat{Y}$, we can answer the question of if there is bias when generating the prediction model.

We coin the following terms: **1)** changes in direct effect $(DE(Y) - DE(\hat{Y}))$, **2)** changes in explainable discrimination: $(ED(Y) - ED(\hat{Y}))$, **3)** changes in indirect discrimination

$ID(Y) - ID(\hat{Y})$. By analyzing, these changes we can potentially have another way to rank fairness mitigation algorithms.

There is no direct interconnect between the causal and fairness metrics because in the real world there are confounders that might affect the causal effect between A and Y. Both direct effect and indirect effect might contribute to the bias detected with disparate impact or equal opportunity difference, but since total effect is associated with disparate impact. So, there is a reduction in TE, we know that disparate impact values will move toward to 1 (the model is fair when disparate impact is equal to 1). Since, if $TE(\hat{Y}) = 0$ then

$$TE(\hat{Y}) = DE(\hat{Y}) + IE(\hat{Y})$$
$$0 = DE(\hat{Y}) + IE(\hat{Y})$$
$$DE(\hat{Y}) = IE(\hat{Y})$$

and $TE(\hat{Y}) = 0$ potentially can be achieve in two way. One way is that $DE(\hat{Y})$ and $IE(\hat{Y})$ are both 0. This means that there is no causal connection from A to Y. Then, potentially any mediator among the causal path is not used for prediction. Another path is when $DE(\hat{Y})$ and $IE(\hat{Y})$ are offsetting each other. Then, we need to conduct a secondary analysis to see if there $IE(\hat{Y})$ consists of explainable discrimination or indirect discriminating.

3.4 Github link and packages used

Our code is available on GitHub. Causal effects are estimated using the R PATHS package (Zhou and Yamamoto (2020)). We used the statistical fairness metrics and mitigation algorithms from Bellamy et al. (2019).

4. Experiments

4.1 Demonstrating the calculation of casual metrics through UC Berkeley 1973 Graduate Admission dataset

One of the classical examples to demonstrate the problems with the statistical fairness notion is Simpson's paradox, which can be summarized by the formulation below:

$$P(Y|A) < P(Y|\neg A)$$
$$P(Y|A, Z = z) < P(Y|\neg A, Z = z)$$

where A is the sensitive attribute, Y is the target variable and Z is other random variable of the dataset. In colloquial terms, this means that although there is bias detected by conditioning on the sensitive attribute, if we condition on the joint of the sensitive attribute and another random variable, this bias can be potentially explained away.

The UC Berkeley 1973 admission data presents an example of Simpson's paradox, where it appears that the overall admission of the university is biased towards accepting more male students versus female students. However, conditioning on the acceptance rate by individual departments, we see that the female acceptance rate is higher across all departments (Table 3).

Therefore, many studies have suggested that the admission rate difference is a result of female students choosing to apply to departments with lower acceptance rates. While this argument might seem feasible and valid, we cannot accept it to be true without understanding the causal effects of the model. The argument can be easily debunked by assuming that

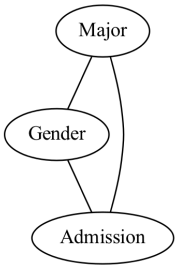

Figure 1: Causal discovery on the Berkeley admission dataset (PC algorithm)

Statistical Metric		**Score**
Disparate Impact	$\frac{P(A=F\|Y=1)}{P(A=M\|Y=1)}$	0.78
Causal Metrics		
Direct Effect	$A \to Y$	0.09 $\pm$.01
Indirect Effect	$A \to M \to Y$	0.006 $\pm$.001

Table 1: Fairness evaluation of admission data.

there is a potential confounder (e.g applicant's aptitude) that affects a student's choice of the department and also the admission outcome, and by conditioning on this confounder, we still discover direct effect of the gender on the admission outcome.

Feeding the dataset through our pipeline, we estimated the direct effect and the indirect effect of the sensitive attribute. The results of the casual discovery algorithm are shown in Figure 1. Information of the direct effect and the indirect effect via the departments are summarized in Table 1.

Analyzing the results from the experiment, we cannot reject the hypothesis that college admission is not determined by gender, since through causal discovery, we still see a direct path from gender to admission. Also, looking at the causal effects, the direct effect is still dominating with of value of 0.9, while the indirect effect is 0.006. That said, this process of computing the statistical fairness metrics along with the causal based fairness metrics is a more comprehensive approach to understanding and mitigating the fairness problem.

4.2 Analyzing the effect of fairness mitigation algorithm on causal and statistical fairness metric through the UCI Adult and ASCIncome datasets

The structural causal model (SCM) for the UCI Adult dataset obtained using the PC algorithm is shown in Figure 2. As done by Binkyt.e-Sadauskien.e et al. (2022), we first introduced domain knowledge in the form of a temporal partial order by defining three tiers. In the first tier, we included `age` and `sex`. In the second tier, `education` and `marital status`, and in the last third tier, `working class`, `number of working hours`, and `income`. Upon inspection, the SCM appears reasonable. There are three paths from the sensitive attribute to the target variable, `income`: 1) The direct path (`sex` $\to$ `income`) 2) the discriminative, indirect path (`sex` $\to$ `martial status` $\to$ `income`) and 3) the explainable, indirect path (`sex` $\to$ `education` $\to$ `income`). The second path is considered discriminating, since marital status is likely correlated with the sensitive attribute, `sex`. This can be due to the fact that females are more likely to take maternity leave once married.

Figure 4 shows the direct and the indirect caused effect on the target outcome Y. The direct effect of gender on income is 0.181, while indirect effect via the `education` path and `marital status` is 0.005 and 0.008 respectively. From this we can conclude that the direct effect of `sex` on `income` is still dominating over the indirect path and that there is

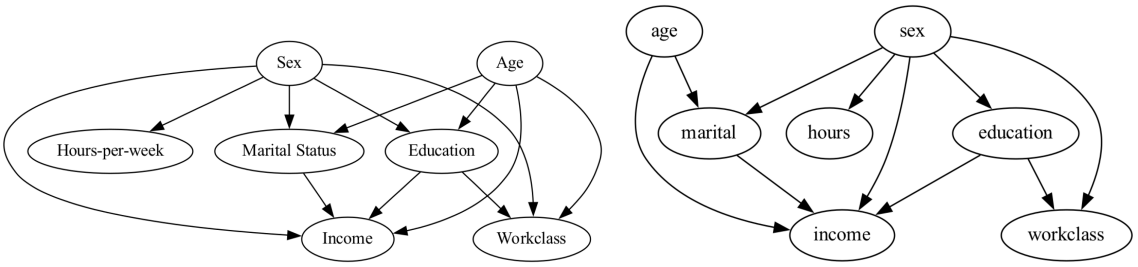

Figure 2: SCM for the UCI Adult and ASCIncome datasets using PC Algorithm

no discrimination via `marital status`. Just by looking at this result, it is mostly likely that disparate impact will flag this model as biased. After training a ERM algorithm, we do observe that disparate impact and equalized odds different show that there is a bias.

After applying the adversarial debiasing algorithm, we observe changes in the causal effect (Figure 4) and the statistical fairness metrics (Figure 3). We can see that both direct effect and total effect greatly decreased and the indirect, explainable effect via `education` largely increased from 0.005 to 0.017. With the reweighting algorithm, we observed the same increase in the disparate impact and equal opportunity difference, but the direct effect of reweighting algorithm decreased as much as with the adversarial debiasing algorithm. With the optimizing preprocessing algorithm, most of the direct effect are pushed towards the indirect effect explained by `education`. After using the rejection option algorithm, we can see that the total effect decreased, but the indirect, discriminative effect via `marital status` is much more higher, which can be problematic since this fairness mitigation method might be introducing other biases to the classifier, increasing the discriminative bias. One key finding is that all four fairness mitigation algorithms were able to address the bias initially present in the data, and while the adversary debiasing algorithm resulted in the largest mitigation, the rejection option algorithm was the most successful in reducing direct effect, which was the greatest source of bias initially.

We conducted the same experiment with the ASCIncome dataset. The results are reported in (Figures 6, 7 and Table 5). Discovering and computing time was much higher due to the large number of samples in this dataset. This dataset also discovered biases through the disparate impact and equal opportunity metrics. All four fairness mitigation algorithms were able to reduce the bias, but only the rejection option algorithm showed a decrease in the direct effect. This is likely due to the larger amount of data or due to problem with implementing the algorithm for measuring causal effects. We are actively looking into this issue.

5. Conclusion

In this paper, we developed a casual fairness pipeline for observational data in an effort to marry statistical and causal fairness metrics to yield a more comprehensive approach. This pipeline can be applied to analyze classification outcomes and give insight into the effect of statistical fairness mitigation algorithms. Statistical fairness metrics have its limitations, such as being unable to deal with the Simpson's paradox, but there are also benefits to

Figure 3: Statisical fairness metrics on UCI adults.

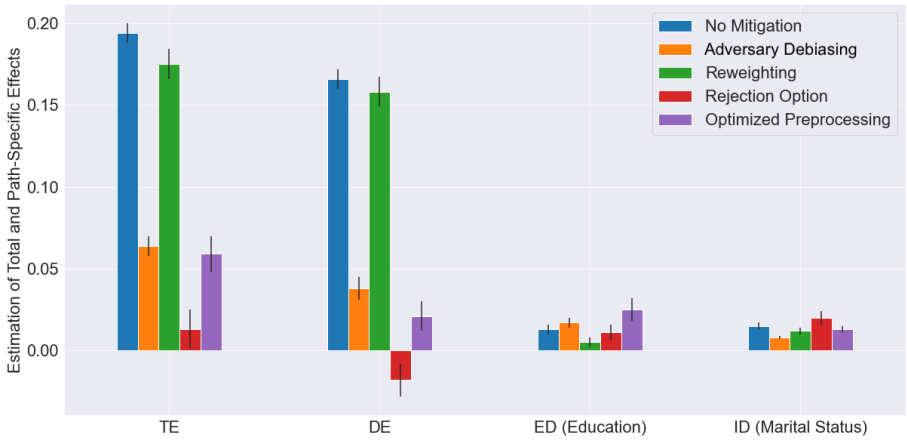

Figure 4: Casual fairness evaluation for UCI Adult.

using these metrics, such as that they can be computed efficiently, and it is more stable as compared to the causal fairness metrics. Causal fairness metrics are dependent on the causal structure model defined, and estimation of causal metrics based on this structure adds extra uncertainty, as the structure is merely hypothetical based independence assumptions on the given population. Through our experiments, we showed that casual fairness metrics can give insight into the specific cause of the bias in a classifier by identifying the path-specific indirect effects. We believe that causal fairness metrics and statistical fairness metrics should be used in combination, especially when multiple fairness mitigation methods obtain the same level of statistical fairness mitigation.

Appendix A.

Feature	Description	Type
AGEP	Age	Numerical
COW	Class of worker	Categorical
SCHL	Educational attainment	Categorical
MAR	Marital status	Categorical
WKHP	Usual hours worked per week past 12 months	Numerical
SEX	Sex	Categorical
PINCP	Total person's income	Categorical

Table 2: ACSIncome Feature Summary

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

Table 3: UC Berkeley 1973 admission rate by department

	Adversary Debiasing	**Reweighting**	**Rejection Option**	**Optimized Preprocessing**
$TE(y) - TE(\hat{y})$	-0.019	-0.003	**0.108**	-0.038
$DE(y) - DE(\hat{y})$	-0.019	-0.001	**0.106**	-0.040
$ED(y) - ED(\hat{y})$	0.002	0.000	0.003	**0.000**
$ID(y) - ID(\hat{y})$	**-0.003**	-0.001	-0.001	**0.003**
Equal Opportunity Improvement	**0.234**	0.381	**0.397**	0.144
Disparate Impact Improvement	**0.260**	0.565	**0.591**	0.415

Table 4: Summarizing of the changes in the causal and statistical fairness metrics for the UCI dataset

	Adversary Debiasing	Reweighting	Rejection Option	Optimized Preprocessing
$TE(y) - TE(\hat{y})$	-0.019	-0.003	**0.108**	-0.038
$DE(y) - DE(\hat{y})$	-0.019	-0.001	**0.106**	-0.040
$ED(y) - ED(\hat{y})$	0.002	0.000	0.003	0.000
$ID(y) - ID(\hat{y})$	-0.003	-0.001	-0.001	**0.003**
Equal Opportunity Improvement	0.234	0.381	**0.397**	0.144
Disparate Impact Improvement	0.260	0.565	**0.591**	0.415

Table 5: Summarizing of the changes in the causal and statistical fairness metrics for the ACSIncome dataset

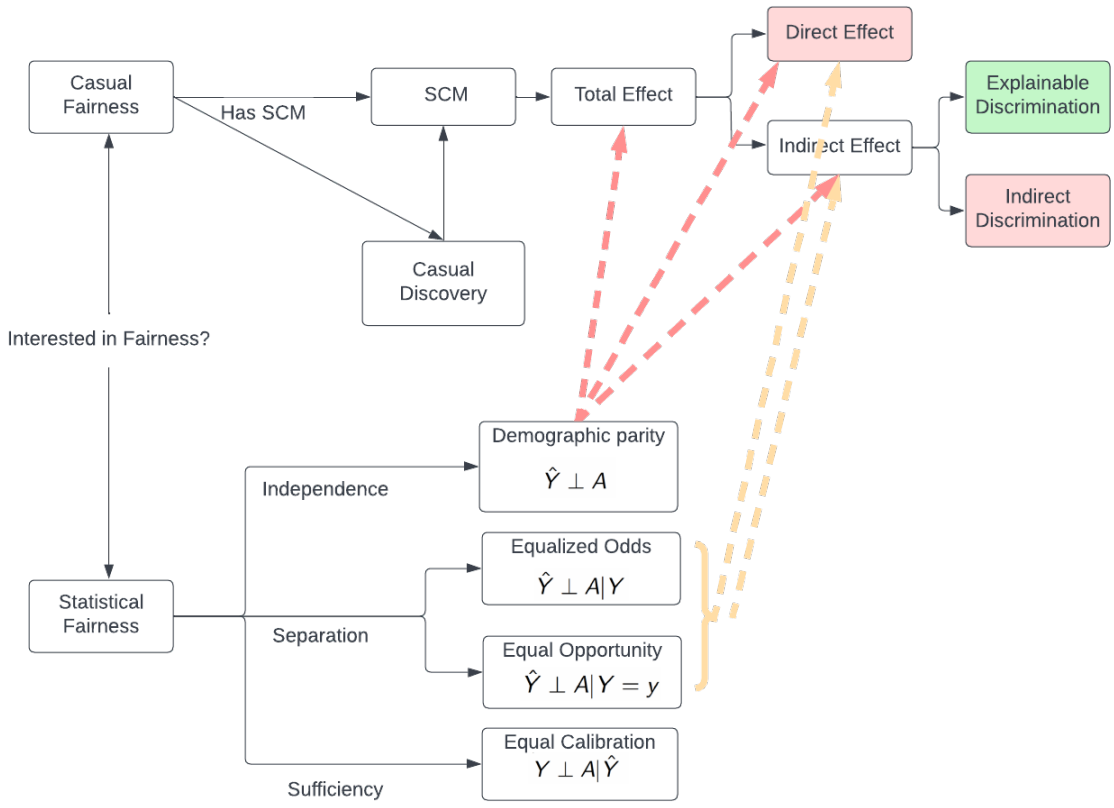

Figure 5: Fairness Comparison Pipeline

Figure 6: Statisical fairness metrics on ASCIncome

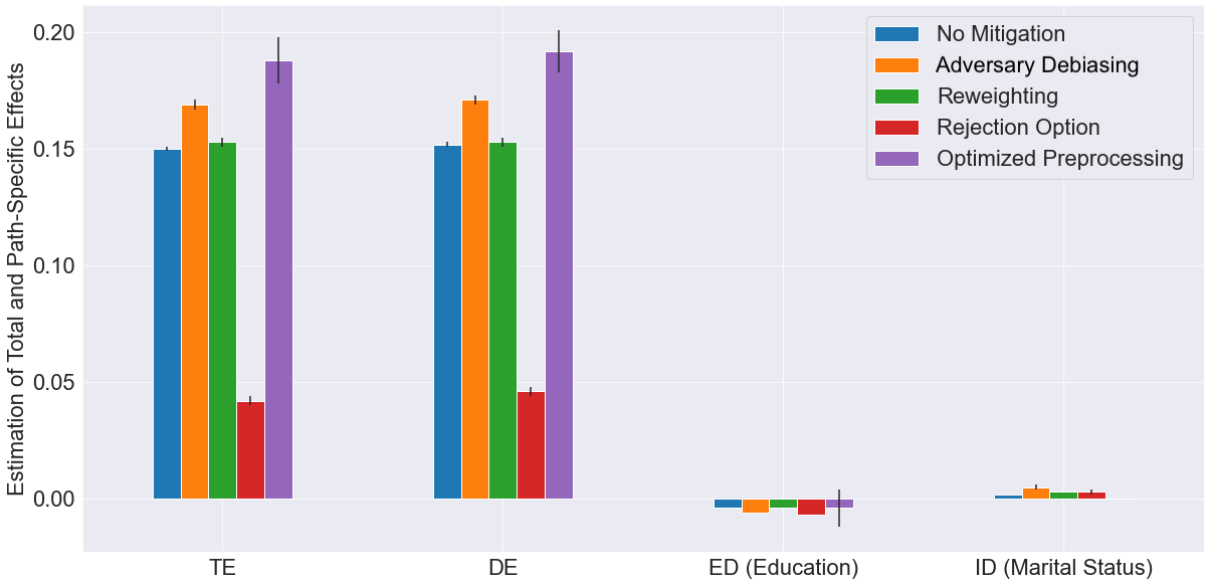

Figure 7: Casual fairness evaluation for ASCIncome

References

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016. ISSN 00081221. URL `http://www.jstor.org/stable/24758720`.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Ruta Binkyt.e-Sadauskien.e, Karima Makhlouf, Carlos Pinz'on, Sami Zhioua, and Catuscia Palamidessi. Causal discovery for fairness. *ArXiv*, abs/2206.06685, 2022.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *ArXiv*, abs/2108.04884, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *ArXiv*, abs/1104.3913, 2012.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. 2016.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *ArXiv*, abs/1610.02413, 2016.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *ArXiv*, abs/1610.09559, 2016.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

R. Kohavi and B. Becker. Uci adult data set. *UCI Machine Learning Repository*, 1996.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *ArXiv*, abs/1703.06856, 2017.

Blake Lemoine, Brian Zhang, and M Mitchell. Mitigating unwanted biases with adversarial learning. 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2016.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *CoRR*, abs/2010.09553, 2020a. URL `https://arxiv.org/abs/2010.09553`.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *ArXiv*, abs/2010.09553, 2020b.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.*, 58:102642, 2021.

Cathy O'Neil. Weapons of math destruction: How big data increases inequality and threatens democracy. 2016.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, 2001. ISBN 1558608001.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Steven Ruggles. Integrated public use microdata series. *Encyclopedia of Gerontology and Population Aging*, 2021.

English Simpson. The interpretation of interaction in contingency tables. *Journal of the royal statistical society series b-methodological*, 13:238–241, 1951.

Richard S. Zemel, Ledell Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

Xiang Zhou and Teppei Yamamoto. Tracing causal paths from experimental and observational data. 2020.